



מדינת ישראל  
STATE OF ISRAEL

Ministry of Justice  
Patent Office

משרד המשפטים  
לשכת הפטנטים

This is to certify that annexed  
hereto is a true copy of the  
documents as originally  
submitted with the patent  
application of which  
the details are specified on the  
annex.

זאת לתעודה כי רצופים  
בזה העתקים נכונים של  
המסמכים שהופקדו  
לכתחילה עם הבקשה  
לפטנט לפי הפרטים  
הרשומים בעמוד הראשון  
של הנספח.

15-08-00  
This is to certify that  
the copy is a true copy of the  
documents as originally  
submitted with the patent  
application of which  
the details are specified on the  
annex.  
Commissioner of Patents

נאשר  
Certified

BEST AVAILABLE COPY

לשימוש הלשכה  
For Office Use

מספר:  
Number  
119183

תאריך:  
Date  
02-09-1996  
הוקדם/נדחה  
Ante/Post-dated

חוק הפטנטים תש"ז - 1967  
PATENT LAW, 5727-1967

בקשה לפטנט  
Application for Patent

אני, (שם המבקש, מעני ולגבי גוף מאוגד - מקום התאגדותו)  
(Name and address of applicant, and in case of body corporate-place of incorporation)

חיים צבי מלמן, חגי 3 כפר סבא

נוחי צותיל, בודוכוב 45 הרצליה

היותנו המציאים

בעל אמצאה מכח  
Owner, by virtue of

שמה הוא  
of an invention the title of which is

(בעברית)  
(Hebrew)

שיטה ומכשיר לחיפוש ואיחזור מידע מבוזר

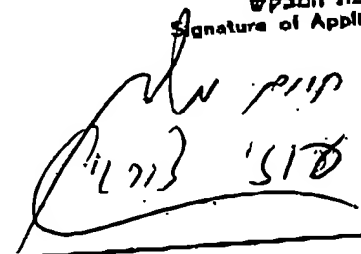
Method and apparatus for search and retrieval of distributed information

(באנגלית)  
(English)

מבקש בזאת כי ינתן לי עליה פטנט

heraby apply for a patent to be granted to me in respect thereof.

• דרישה ריץ קרימה  
Priority Claim

• בקשה חלוקה - Application of Division	• בקשה פטנט מוסף - Application for Patent Addition	• דרישה ריץ קרימה Priority Claim		
מבקשת פטנט from Application	• לבקשה/לפטנט to Patent/Appl.	מספר/סימן Number/Mark	תאריך Date	מדינת האגוד Convention Country
No. _____ dated _____	No. _____ dated _____			
• יסוי נח: כללי / מיוחד - רצוף בזה / עוד יוגש P.O.A.: general/individual-attached/to be filed later- הוגש בענק filed in case _____				
המען למסירת פטנטים בישראל Address for Service in Israel חיים מלמן חגי 3 כפר-סבא 44335				
חתימת המבקש Signature of Applicant 		1996 שנת 26 בחודש _____ of the year _____		

לשימוש הלשכה  
For Office Use

סדקס זה כשרוא מוטבע בחותם לשכת הפטנטים ומשולם במספר ובתאריך ההגשה, העו אישור להגשת הבקשה שפרטיה רשומים לעיל.  
This from, impressed with the Seal of the Patent Office and indicating the number and date of filing, certifies the filing of the application the particulars of which are set out above.

מחק את המחקר  
Delete whatever is inapplicable

שיטה ומכשיר לחיפוש ואיחזור מידע מבוזר

METHOD AND APPARATUS FOR DISTRIBUTED INFORMATION SEARCH  
AND RETRIEVAL

## FIELD OF THE INVENTION

The present invention is related to search and retrieval of data and information distributed among independent databases, such as Internet servers.

## BACKGROUND OF THE INVENTION

During the last decade text retrieval has been developed to a large extent. In the last few years the volume of digital text files has increased exponentially, making full text indexing and retrieval a major branch in database management.

Today main stream for handling text files is based on client-server architecture. Solutions are based on the concept that both the database, the development and the indexing of the database, as well as the search and retrieval tools and application are under the full control of the MIS (Management Information Systems) team.

The evolution of the Internet has generated a new model for document databases and retrieval. Web servers, indexing robots, search engines and web navigators has turned the web to a huge, constantly changing document resource. The concept of well controlled databases, search and retrieval applications is no longer valid. Databases distributed among the servers of the Internet are "ever changing, individual entities".

The data and information located in such servers is selected and managed by a local team to respond for some local requirements, following local rules, ignoring the relation of the content of that database to the content of other databases in other Internet servers. Some servers concentrate on entertainment information, others on commercial subjects.

Search engines in different servers obey different search rules. Some will accept a high level boolean query, others are limited to the "OR" boolean operator only. Search engines are limited to local server search.

New databases emerge and other are eliminated in an increasing rate, requiring constantly growing investments in updating of resources list.

Many people from the organization level to home level, for professional through business to hobby needs, spend allot of time to find useful documents in the web. Present search methods are relatively primitive and time consuming. Many times the results are disappointing. This is due to few reasons:

1. Information in the Internet is spread all over the net in a chaotic manner. In most cases a user would not know which Internet server/host contains the information they need. As a result, they may be looking for information in a server that is not an "expert" in that type of information.
2. The user may not be aware of new information sources as well as of the elimination of existing sources.
3. In many servers the search tools are not sophisticated enough, resulting in low level data filtering.
4. One need to learn how to work effectively with a variety of search tools, associated with a variety of information providing servers, the search tools vary in search capabilities.
5. Search through different servers is not correlated. This results in unnecessary multiple retrieval of the same data from different servers.
6. The relevance ranking of data from various servers is based on unrelated ranking algorithms, resulting in relevance ranking confusion among data from different sources.
7. Present Web browsers provide manual hypertext search that occasionally generate closed loop - repetitive returns to unwanted locations, wasting the time of the user and generate unnecessary load on the net and servers.

In U.S. Pat. No. 4,774,655 Kollin offers a system that can be communicated by many users. The system contains a classified subject list. The end of each branch provides a reference to databases related to the subject indicated. According to Kollin, the maintenance of the resources list is made by a "...skilled reference librarian..." (column 2 lines 18-26). The user must first select the specific database directly or by selecting a subject and then place the search query. Kollin also offer a query translation from the user syntax to the server syntax. However, if the logical capabilities of search engine of database are incompatible with the query (such as not having a boolean NOT operator that the query contains). Kollin suggest an error message to the user and reformulating the query (column 12 lines 54-58). After receiving the data from each database the user browses through the data to evaluate it and decide if a search in another database is required. Kollin offers methods for reducing cost of using commercial databases by connecting to an intermediate "agent" system that provides services at a lower cost and reduces the usage time/cost of the databases themselves. In the Internet, as is today, the issue of cost for using databases relatively neglectable. A huge amount of free databases is available at the low cost of Internet service provision. The user time spent on searching for data has become the main issue.

Kollin, in spite of providing few solutions to overcome the user limited knowledge in databases, fail to provide solutions to save data handling time and remove dependency on a database librarian:

the present invention solves the problems described above by providing a new method and apparatus comprising:

1. A method for selecting databases without subject selection by the user.
2. A method for automatic adjustment of a query to a search engine of a database that has incompatible set of logical operation. This is done with out losing the full scope of the original query logical relations specifications. It including a databases of which the query logical capabilities are unknown
3. Automatic procedure for traversing multiple databases, particularly databases which are not on a resources list.
4. A method for avoiding multiple retrieval of the same data from different sources, consolidating the data coming from many sources and providing a common basis ranking.
5. A method for consolidation and classification of data retrieved from plurality of sources
6. A method for automatic construction and updating of a database resource list without the need for an expert. The source list is optimized to the specific user area of interests.
7. A method for generating a local database with a high relevance to the user fields of interest.

## SUMMARY OF THE INVENTION

It is the objective of the present invention is to provide search and retrieval tools, particularly useful for document retrieval from distributed and unrelated data bases. In particular, the suggested method is designed to take advantage of the nature of the Internet as an information source constructed from many distributed and unrelated databases. It includes features that save most of the user's time spent today on information searching and provide better search results then available with the present art.

Due to the close relation to the Internet, terms of the Internet will be used without special notation, as these terms are familiar to those skilled in the art.

In accordance with the present invention, a search and retrieval method and apparatus are provided, comprising at least some of the following:

### 1. Client Station

- Client Reference Database
  - Client subgroup of classified subjects
  - Client statistical data on hosts (such as Internet hosts), including at least one of:
    - List of names of hosts approached in the past with rate of success and failure accumulated for these hosts.
    - Number of returns to a host for documents retrieval.
    - Host ranking that represent the value of a host to the client, based on past success statistics.
- A Query Module having an interface for submitting a search query including at least some of:
  - Boolean query editing interface
  - A reference to at least one Internet URL (or equivalent in a similar system)
  - Selectable classified subjects
- Query Interpreter for query processing
- Document analyzing module with algorithms including elements such as Boolean relations, statistical relations and thesaurus
- Data base updating procedures

### 2. Local/Client Server

- Local/Client Server database
  - Local subgroup of classified subjects list (consolidation of the data bases from the various clients, connected to this server and represent an organization)
  - Local hosts statistical data
    - Names of hosts approached in the past with rate of success and failure
    - Number of returns to a host for documents retrieval
  - Directory for approved documents
  - Popular documents (to reduce Internet load)
  - Automatic maintenance module (updating from search results)
  - A list of Internet information providers
    - Characterization of search engine at remote server (information provider)
    - Ranking according to relevance to a classified subjects list (graded levels of expertise for each provider)
    - Automatic update procedure (updating from another database)
- Query interpreter
  - Including complexity reduction of a boolean expression
- Local search module

3. A group of unrelated remote servers - such as the Internet with it's hosts servers
  - Each such remote server is equipped with a specific index and a specific search engine to respond to a query from the client and engage it's search engine to find relevant information. The relevant information is then sent to the client.

Main steps of a simplified query procedure are described below, demonstrating that the user is not required to be familiar with remote servers that may have the requested data, nor has he to be familiar or limited to the sophistication level of the search engines of remote servers:

1. Generate and submit a boolean query such as *(wagon or vehicle) and (rail) not (road)*
2. The Boolean expression is "broken" to it's component key words: *wagon vehicle, rail, road.*
3. The key words are used with a local reference database to search for hosts that are relevant to these key words.. A list of relevant hosts is made, the hosts are ranked according to their relevance to the key words.
4. The characterization of the search engine of the first host is retrieved from the local reference database. In this example it is capable of full boolean query. The original query is reorganized in the format expected by the search engine and is submitted.
5. The results from the first host are received and filtered and ranked at the client, using the client filtering and ranking modules. They may include titles, summaries and documents.
6. If the number of findings is not sufficient, the program proceed to the second host in the hosts relevance list.
7. The characterization of the search engine of the second host is retrieved from the local reference database. In this example it is only capable of a simple query, such as *(wagon or vehicle or rail or road)*. The original query is reorganized in the format expected by the search engine and is submitted to it.  
To avoid at least some of the information containing *road*, the Query Interpreter omits the word *road*, arranges the query on *wagon or vehicle or rail* in the format suitable for the second host and submits the query.
8. The results from the second host are received and stored at the client. Then these results are filtered and ranked again using the client search that is capable of performing the filtering with the original boolean expression *(wagon or vehicle) and (rail) not (road)*. Results that do not satisfy the boolean expression are deleted.
9. If the number of results acquired from both hosts is sufficient, both groups of acquired results are consolidated and presented to the user according to the client relevance criteria.

The Internet was selected as a an excellent example for distributed and unrelated information offering servers. Although some of the features of the present invention takes advantage of special properties that are provided on the Internet (such as hypertext searching with URL references), this does not limit the invention to the Internet. The invention is applicable, in different degrees, to any information retrieval net that comprises at least two data bases.

Parts of the detailed description of the invention is written in reference to the World Wide Web environment of the Internet (known also as the Web, WWW or W3). This is done in purpose of simplifying the description. It will be appreciated, by those skilled in the art, that

the implementation in other environments (such as Wais - Wide Area Information Servers, FTP - File Transfer Protocol or even non Internet distributed-unrelated databases) is analogues.

Also, in general, document search techniques are not described in details in this document as they are familiar to those skilled in the art. Reference to such techniques are available through vendors of full-text retrieval systems such as Varsity of Mountain View, California, PLS of Rockville, Maryland and Fulcrum Technologies Inc. of Ottawa, Canada.

The query interpreter may be also used in specific cases of a single database, having it's specific search engine that is not familiar to the user. The user may then use his search interface, having the interpreter adjusting the query to the specific search engine. After the results are provided, the user may re-filter and re-rank the results using his local search and ranking application.

The invention will be described hereinafter by way of example with reference to the accompanying drawings, wherein.

Figure 1 is an overview of the Internet as a net of distributed, unrelated information providing servers.

Figure 2 is drawings of components arranged according to one embodiment of the invention, useful in performing a search session.

Figure 3 provides a flow diagram that is useful in one embodiment of the invention.

Figure 4 provides more details of Figure 3, relating to the URL/hypertext based search.

Flow diagrams are provided for illustration purposes and the scope of the invention is not limited to what is shown in these diagrams.

## DETAILED DESCRIPTION OF THE INVENTION

A general example for a network that connects many sources of information that are unrelated to each other is the Internet. An overview of this network is provided in Figure 1. The Internet backbone (102) comprising fast communication lines. The backbone can be approached through computers called "gateway" (104) that control the communication between the backbone and the networks (106) attached to the gateway. Networks may comprise servers (that may also act as local gateways) such as (108), (110) and (112) and clients such as (114), (116) and (118). Internet servers that allow clients interact with the server and it's content are called often "hosts". For more detailed description of the Internet structure, communication protocols such as TCP/IP and other related topics, see

- "Using the Internet 2ND edition" by Mary Ann Pike, published by Que Corporation USA
- "Internet Architecture Board" (IAB) publications, of which many are in [ftp.internik.net](http://ftp.internik.net).

In the following embodiment of the invention, the methods and tools enabling the process will be described by following the process steps and describing the tools associated with each step.

Some of the tools that are described in prior art (Client Reference Database) will be described hereinbelow in details to provide a comprehensive description of the present invention.

### 1. Generating a Search Query

Reference is made to Figure 2. A Query Module (QM), 204, is provided on a client station. The properties of this module are independent of the hosts that will be searched. They depend only on the needs and requirements of the user. For example, the module may contain



- and, or, not .boolean relations
- free text expressions
- classified subjects list

The user uses the QM GUI (Graphic User Interface) 202 to compose a query such as (wagon or vehicle) and rail and 'Transtech 96' not road.

The term 'Transtech 96' is a free text element being, for example, a name of a specific transportation exhibition.

The QM and the associated GUI may be more complex and it may be used in a variety of ways.

More search features that may be available in the GUI of the QM:

- Setting the maximum number of results to be retrieved.
- A choice for searching through any Selectable combination of: titles, abstracts, full text, author or any other relevant section of the document.
- A data relevance minimum level is set to indicate the range of relevance that will be accepted.
- A selection of a URL to be used as a starting point for URL/hypertext based search.
- A selection of one or more subject out of a classified subject list.

When the user has finished to compose the query and other search parameters, he selects a "submit" menu item to forward the query to the Query Interpreter, 206.

## 2. The Client Reference Database

The concept of a Client Reference Database (CRD) is shown by Koliin, U.S. Pat. 4,774,655. The concept is described here in details to provide the basis to the description of embodiments of the present invention, among them are:

- Automatic generation and maintenance of a CRD
- Automatic selection of hosts for search from the CRD.

As mentioned hereinabove, the user may also specify the subjects that his query is associated with. For example, he may select "Electrical vehicles" from the following illustration of classified subject list:

```

Transportation
  Land
    Power sources
      Liquid fuel vehicles
      Electrical vehicles
        Cars
          Personal
          Family
          Trucks
        Trains
        Gas driven vehicles
        Road vehicles
        Off track vehicles
        Rails
        Statistical data
    Air
    Sea
  
```

It is a particular embodiments of the invention that the user may select more than one subject and he also may decide not to select any subject at all. It is possible to ignore the boolean option and use only the classified subject list.

The CRD, 208 in Figure 2, contains at least one of the following elements:

- Classified subject list
- Hosts names that are relevant to the CRD
- Relevance ranking of hosts to the subjects with which they are associated
- Host characterization:
  - List of available services from each host.
  - Parameters required for working with that host and its services.
  - Information on using the host search engine.

Other useful utilities may be provided to expand the CRD such as:

- Thesaurus
- Speller

It would be appreciated that the simplest form of a useful CRD is one that contains only a list of information sources. In the Internet example, such a CRD will take the form of a list of URLs. The next level will include information on query placement specifications.

The classified subject list of the CRD may be constructed as any other classified subject list (such as Yellow Pages) with the addition that each subject contains a list of hosts and relevance ranking list. An example format may be of the form:

<subject> [<host name>/<relevance ranking>, ...]

<subject>:

A line in the classified subject list.

<host name>: The full address of the host and the service type. For example <http://www.uspto.gov>. A specific host may be mentioned more than once for each of the relevant services available from that host (assuming no complete overlap of the information available through different services).

In the description, host name and service will be indicated in the form "host123".

<relevance ranking>:

This indicates the relevance of the specific host to the subject. Relevance ranking of a host is based on ranking algorithms such as accounting how many times the subject is indicated in that host. In this example relevance ranking will assume values 1 to 10, 10 being highest relevance.

A host may be listed more than once, with different subjects. For each such subject the host may have a different ranking. It will be appreciated that for each such host association with a subject, the host can be regarded as different hosts.

Since a host may be listed more than once it is useful to maintain a separate list of the hosts, containing the host information which is identical for all said listings. That is information such as communication protocols, parameters and services available from that host.

Following the example of classified subject list set above, some of the subjects are shown hereinbelow with the associated hosts names and relevance ranking :

Transportation [host867/10, host1/7, host83/7, host12/5, host155/5,  
host654/3, host34/1, host935/1]

Land [host83/10, host867/8, host1/7, host12/5, host155/5, host654/3]  
Power sources [host83/10, host867/8, host1/7, host155/5,  
host654/3]

Liquid fuel vehicles

Electrical vehicles [host83/10, host1/10, host867/8,  
host155/5]

Cars [host83/10, host1/10, host155/5, host867/3]

Trains

Gas driven vehicles

It is a particular embodiment of the present invention to associate hosts to all levels of the CRD, thereby, saving the user the need to make a multiple level subject selection, throughout the end of the subjects branch. It is also a particular embodiment of the present invention, that a host may listed and unlisted along the different levels, having different ranking at each level, as described hereinbelow.

Looking at the subject "Transportation", a list of 8 hosts is associated with it. These are at least some the hosts, out of a larger list of accessible hosts, that are ranked 1 or higher using the ranking process. Using this list, *host867* would be the first host to be searched for information. Then only if there is not enough information found, *host1* will be searched, and so on.

If we look at the subject "Land", which is actually "Transportation, Land" we can see that for that category, *hosts34* and *host935* are not relevant anymore. Being ranked under 1 they are deleted from the list. Also, the rank of *host83* has increased from 7 to 10 while *host867* decreased to 8.

Similar structure is applied throughout the classified subject list. The selection of "Electrical vehicles" which is actually a selection of "Transportation, Land, Power sources, Electrical vehicles" is associated with a relative narrow and focused group of hosts to search in.

It will be appreciated that selecting "Electrical vehicle" under a different category such as "Environmental, Pollution, Transportation, Electrical vehicles" would not necessarily have the same hosts list associated with it. The ranking and the hosts names may be different

Relevant hosts may be located using hypertext-URL search, robots (see for example <http://info.webcawler.com/mak/projects/robots/robots.html>), manual browsing, published information in professional literature ect. Updating of the CRD may be automatic or manual as described hereinbelow.

It will be appreciated that updating of CRD may be done at least at one specific host, supported by a team specializing in studying the Internet and upgrading such a CRD. The client may keep updated by receiving updated CRD from such a host on a regular base by an e-mail notification for update. The client may also initiate the update following a routine check such as:

Check last update of the host CRD. If update has been made and at least 2 weeks has passed from the previous update - download the new CRD to your station.

### 3. Search Query Interpretation

The search query is submitted to the Query Interpreter (QI), 206 of Figure 2. In the present example it contains:

- The boolean expression:  
(wagon or vehicle) and (rail) and ("Transtech 96") not (road).
- The selected subject (or subjects):  
"Transportation, Land, Power sources, Electrical vehicles"  
with it's associates hosts list:  
[host83/10, host1/10, host867/8, host155/5]

Another preferred embodiments of the invention that either do not use a subject selection or use only subject selection will be described hereinbelow.

In order to retrieve documents from a host, the QI generates modified sets of search queries, reorganized in a format suitable to that host. Each one of the modified queries is submitted to the Net Communication Module (NCM), 210, a software module that is designed to send the query to the selected host (and receive the response from that host). This is shown in Figure 2 as a communication path of NCM 210 to net 212 which is connected to many hosts and servers 214.

The NCM will not be described in details here. There are equivalent software elements in Internet browsing software using CGI (Common Gateway Interface). One such example is Navigator from Netscape Communications Corporation of Mount View, California, another example is Mosaic from the University of Illinois. Also a variety of development tools are available to generate NCM applications such as The Internet Application Framework from Netscape Communications Corporation and developers tools from Spyglass Inc. of Naperville, Illinois.

It is appreciated that the development of such a module is a common task for those skilled in the art.

### 3.1 Search Process

Since, as demonstrated in the present example, the user specified a subject for the search process, this will dominant the first phase of search. In this example, using the classified subjects list ranking of hosts, *host83* will be the first choice for search.

In a preferred embodiment of the invention, *host83* enables processing only a single term with it's search engine. That is, the boolean relations such as AND, OR and NOT are unrecognizable by the search engine of that host. To utilize such a host without reducing the benefits of a complex boolean query, the QI submits 5 separate search queries, one query for each of the terms:

*wagon, vehicle, rail, Transtech* and 96.

The term *Transtech 96* is split to *Transtech* and 96 since *host 83* does not support a submission of anything but a single word.

The term *road* is not useful at this stage as it may result in documents containing only the word *road*, documents in which the user is definitely not interested. The word *road* is identified as such by the preceding boolean relation *not*.

It will be appreciated that this method may be used also when the search boolean specifications of the host are not available to the QI.

A common denominator for all search engine is the search of a single term. In different databases, however, the format for submitting the term may be different. This is problem is limited if the search is directed to all hosts capable of handling HTML and/or other widely used format.

Each query will invoke a search process in *host83*, returning a list of titles, abstracts or any other information requested from that host. Full text can be automatically retrieved by subsequent communications with the host. The text is referred to by the URLs that come with the data from *host83* in the preceding search process. Most Internet search engines provide a list of titles of relevant documents and locations. The URLs are associated with the titles in that list. This can enable a search on the level of full-text without the user interference. The number of data elements will be limited by the relevance level (set by the user during the query composition) or by an internal default limit of the NCM.

When a predetermined amount of information has been retrieved by sending the 5 queries to *host83*, the Client Search Engine (CSE) is set to work (216). The CSE search in the data using the complete boolean expression (*wagon or vehicle*) and (*rail*) and (*Transtech 96*) *not* (*road*). This is possible since the CSE does not have the limitations that the search engine of *host83* may have.

At the end of this process, only documents that satisfy the original boolean expression are identified as relevant.

Reference to the already retrieved documents is maintained. This is used to avoid repeatable processing of such documents as they may be available from other hosts. The reference may be maintained in a special temporary log file that includes useful data such as the document URL, the title and the abstract. For example, before retrieving a document from a second source, the URL associated with the document is compared to the URLs in the retrieved documents log file. If the URL appears there, the document will not be retrieved.

If the process has not generated the requested amount of information, the process will continue with *host1* which also has a relevance value of 10. In this example, the search engine of *host1* can receive a list of terms. An option to select one of two boolean relations is available: *and* / *or* to indicate the relations between all the terms in the list.

The QI will generate two modified queries:  
First query will be for (*wagon, rail, Transtech, 96*) with the *and* option.  
Second query will be for (*vehicle, rail, Transtech, 96*), also with the *and* option.

After receiving the data from *host1*, the acquired data is searched again using the CSE to disqualify the documents containing *road* and ranking the remaining documents using the CSE ranking algorithms.

If there are not enough results, the process will continue to *host867* having a relevance value of 8. In this example, *host867* has a search engine with full boolean capability. In this case, the QI will generate only one query of the original form: (*wagon or vehicle*) *and rail* *and Transtech 96 not road*. The acquired data may be reviewed by the CSE to enable a common base relevance ranking, based on the CSE relevance algorithm.

At this point of the example, we assume that the desired amount of information has been acquired.

As the automatic search process has been finished the data acquired from the different hosts is consolidated to a single group of search results, having a uniform ranking base. The data is then displayed to the user. The user may brows through this data and qualify or disqualify data elements. If by the end of this process the user is left with too little data, he can restart the automatic search process to acquire more data. The process will restart at the point it stopped for displaying the results.

The process described hereinabove will be described by the following steps in reference to Figure 3:

1. The user composes a query, including steps such as the selection of one or more items from a classified subjects list, number of maximum items to retrieve or any other relevant parameter (302).
2. The query and related parameters are transferred to the Query Interpreter (206 of Figure 2), where it is analyzed (304).
3. The first host is selected for search (306).
4. The Query Interpreter reads the host parameters from the Client Reference Database (CRD), such as available boolean complexity. If this is not available, default is used; no boolean capabilities (308).
5. The query is converted to a set of host-compatible-queries, each host-compatible-query is suitable to the requirements of said host, the complete group of host-compatible-queries encloses the scope of the original query (310).
6. Each of the host-compatible-queries is submitted to the host (312) and the data received from the host is accumulated (314). This is done using the NCM (210 of Figure 2).

7. The accumulated data from the set of host-compatible-queries is filtered and ranked (316) using the CSE (216 of Figure 2).
8. An evaluation is made if the results of the search are satisfying (318). The evaluation algorithm may weight parameters such as the number of items and the rank of these items but it may also simply compare the number of items to the predetermined value set by the user.  
If the results are satisfying, they are displayed to the user (320).  
It will be appreciated that partial results may also be displayed and updated many times before the final display.
9. If results from that host are not satisfying and more items are required, then the hosts list is checked for the next host (322). If a next host is available, the next host is selected (306). The process repeats itself until the list of relevant hosts is exhausted.
10. If results are not satisfying and relevant hosts list is exhausted, a URL/hypertext based search can be started (324). One embodiment of the URL/hypertext based automatic search is described hereinbelow, in reference to Figure 4.
11. If results from URL/hypertext based search are satisfying (326), results are displayed (320). Otherwise, a classified subjects list based search (328) can be started.  
One embodiment of the classified subjects list based search is described hereinbelow.  
It will be appreciated that the order of performing different search routines (such as URL based and classified list based) is described in this embodiment in a way of example and it does not limit the scope of the invention.

(CLAIM C: AUTOMATIC URL/HYPertext BASED SEARCH) Reference is made now to Figure 4, describing one embodiment of the present invention: the URL/hypertext based automatic search. The process of URL/hypertext based search is described in the following steps:

1. The documents retrieved in the previous step are used to extract a list of URLs from the URLs associated with them (402). The URLs are ranked according to the relevance of the hypertext segments that refer to the URLs to key words of the original query and according to the rank of the document in which they were found. In many cases the hypertext segment associated with a URL is insufficient for the evaluation and ranking of the URL according to the query. For example the hypertext segment associated with the URL may be simply the Internet address of the URL. Therefore, a good general practice is to evaluate not only the hypertext segment but also the text in the neighborhood of that segment. It may be the sentence containing the hypertext segment, the paragraph or even (as mentioned) the whole document. Each such part of the document may have a different weighted contribution to the ranking of the URL as an information source.  
Also, if the host indicated in the URL is listed in the CRD, the relevance of the host may be compared to the search needs using host ranking in similar subjects as indicated in the CRD.
2. Document is retrieved using the URL of highest rank (404).
3. The document is filtered and ranked (406) using client's local mechanism - CSE.
4. If results are satisfying - such as enough documents has been retrieved and acquired a suitable relevance ranking level (406), they are displayed (320). Otherwise, URL list is checked for more URLs (410). If there are more URLs, the process repeats on retrieving next document (404), until the list of URLs is exhausted.
5. If the list of URLs is exhausted, a decision is made if to proceed to the next level of URLs. The decision may depend on a preset value, defined by the user. The next level of URLs will

be based on URLs from the documents retrieved using the previous URL list. The process of preparing the next level of URLs begins by repeating the sequence from step 1 (402).

This routine continues until the last level of URLs, as determined by the user, is processed.

6. At this stage, a classified subjects list based search (328) may begin.

Classified subjects list based search:

1. Generate a list of new hosts (hosts that were not used before in this search session) by automatic search of the classified subject list, using key words from the query (302). Thesaurus may be used to expand the range of key words (*wagon or vehicle*) to include also words such as *car* and *automobile*. Boolean relations may be configured, automatically or manually, to replace (*wagon or vehicle*) and (*rail*) and (*Transtech 96*) not (*road*) with an expression which suits better the classified subjects list (*wagon or vehicle or rail or Transtech 96*).

Using key words and searching the classified subject list with these key words, will generate a reference to new hosts for search, that were out of range before. Following the example mentioned hereinabove, subject category such as "Environmental, Pollution, Transportation, Electrical vehicles" in the classified list will now be available. The hosts associated with it will be referred to and searched.

2. After the list of new host is prepared, the process continues very similar to the first search phase as shown and described hereinabove in reference to Figure 3, starting at point 306.

It will be appreciated that the process is flexible and it may be configured to include only part of the steps described hereinabove. More steps may be added and the order of the steps may be changed. This may be done by a user selection or it may be programmed in advance.

For example, the process may not include a search by URLs and the user may choose not to select maximum number of documents (then default may be used).

In yet another embodiment of the invention the user does not specify any subject of the classified subject list and thus, does not point at the hosts to search, using his classified subjects list. In such a case, the process may start by automatic generation of the hosts list using the key words and the CRD. This is done by starting the search with a classified subject list based search, as described in steps 1 and 2 hereinabove. After preparing the hosts list, the process repeats the basic pattern of Figure 3, starting at "Select next host" (306).

It will also be appreciated that the generation of hosts list for search may be based on a classified subjects list of any host. Instead of using the classified subjects list of the CRD, one may use a classified subjects list of a reference database that is available at any host and is open for external use. In this method the relevance ranking of associated information resources is generally not available but this method may still provide useful information.

During the performing of the search procedures described above, a host may be suggested for search more than once. It may result, for example, because the host is associated with more than one branch of the classified subjects list, all these branches being relevant for the specific search.

To avoid repetitive search in a host that has already been searched in a given session, a control routine is provided. The control routine keeps track on hosts that has already been searched, such as a log file, and generates a "skip" action whenever that host is targeted again.

In a preferred embodiment of the invention, this control routine keeps track also on URLs. This is particularly important since a URL/hypertext based search may refer to the same documents many times from different paths as well as encounter a closed loop.

#### 4. Searching without a classified subject list

The URL/hypertext based search eliminates the dependency on a CRD as described hereinabove.

In this preferred embodiment of the invention, one or more general purpose databases are listed. Using the Internet example, the list may contain hosts such as:

- InfoSeek Net Search (<http://www.infoseek.com>)
- Yahoo Search Engine (<http://www.yahoo.com>)
- WebCrawler (<http://webcrawler.com>)

One of such hosts may be used as the only (and default) starting point for a URL/hypertext based search. The modified query is submitted to the host. The retrieved documents are filtered locally using the CSE. The URL/hypertext process takes place, using the filtered documents, to generate a ranked list of URLs for further search. The process continues as described hereinabove, for the URL/hypertext based search.

In another embodiment of the invention, the first step of generating the first level of URLs is done successively on all the "general purpose databases" listed. The URLs received from all these databases are consolidated into one group, arranged in priority according to the common relevance ranking. This enable starting with the best URLs of all these databases.

#### 5. Searching without a search query

In this embodiment the user does not specify a query. Instead the user selects a subject from a comprehensive classified subject list.

The subject list may have a reference to hosts - in which case this reference will evoke a CRD based search.

The subject list may not have such a reference - in which case a URL/hypertext based search will be evoked.

In this embodiment the hierarchy of the subjects in the classified list is interpreted to generate a boolean search query. The sample classified subject list will be used to explain the method:

```

Transportation
  Land
    Power sources
      Liquid fuel vehicles
      Electrical vehicles
        Cars
          Personal
          Family
          Tracks
        Trains
        Gas driven vehicles
        Road vehicles
        Off track vehicles
        Rails
        Statistical data
  Air
  Sea
  
```

Selection of Electrical vehicles will generate a query of the form:

*Transportation and Land and (Power w sources) and (Electrical w vehicle).* the w stands for adjacency of words.

A variety of algorithms may be used for composing a query from a selected subject. One such simple algorithm is reflected in the above example:



1. All keywords that constitute a level are specified as adjacent terms enclosed in parentheses.
2. All terms of step 1 are related by an *and* boolean operator.
3. All the terms in the branch of that subject that are leveled prior to the level of the selected subject, and the term of the subject itself, are in the query.

More sophistication may be used by adding truncation operators and using thesaurus to add synonyms with the *or* boolean operator.

#### 6. Reduction of a boolean search expression

Algorithms for reducing complexity of boolean search expression are dependent on boolean operators available in the search expression and those available in the host to be searched. In this example we shall assume an extreme case that can cover practically all *www* information providing sites.

It is assumed that the search query may contain the boolean operators *or*, *and*, *not* and *()*. It is also assumed that the search engine of the host to be searched is capable of processing only a single word in a search session.

The algorithm may consist of the following steps:

1. Make a list of all key words (words that are not boolean expressions).
2. Remove from the list all terms or words preceded by the *not* boolean operator in the original query.
3. Generate a set of query: each query contains only one word from the filtered list.

For example, the query *(wagon or vehicle) and (rail) not (road or track) not (car and race)* will generate the list of following words: *wagon, vehicle, rail*. Each word will be submitted for search in a separate search query. The retrieved documents will be filtered locally, using the complete query expression.

#### 7. Working with CRDs

CRD, Client Reference Database (208) may present a variety of characteristics:

- It can be stored on the client station as a local reference database.
- It can be stored and maintained by specialists at any station, such as a remote Internet host or a corporation server.
- The rank of hosts in the classified subject list can be changed automatically according to the results of a search.
- An interest group may use a shared CRD that is updated automatically by the search results of all group member.
- A new CRD may be automatically generated.

One preferred embodiment of the invention includes the feature of updating a CRD according to the search results. The main parameters associated with such a process are:

- The number of documents that were found at each host.
- The number of documents that were approved for each host.
- The relevance ranking of the documents.
- The number of documents rejected by the user for each host.

An example of updating a CRD is provided by the following steps:

1. All the branches in the classified subject list that are relevant to the query boolean expression are identified, using automatic search process through the classified subject list. Thesaurus may be used to expand the range of the key words so that all the relevant branches are identified (as explained hereinabove, "classified list search").

2. For each such branch get the list of hosts and their ranking.

3. For each such host, calculate a new ranking using a formula:

$$NHR = OHR \times (0.3 \times RF + 0.4 \times (4 \times AF) + 0.3 \times (NH \times RN))$$

where

NHR=New Host Ranking.

OHR= Old Host Ranking.

RF= [Average(Rank i)]/OHR

RF, the Rank Factor is the average of the rankings of the documents received from that host divided by the OHR - Old Host Rank of this host. Therefore, if the average ranking of the documents is higher then that of the host, this will contribute to increase the ranking of the host.

AF= (number of documents approved)/(total number of documents)

AF, Approval Factor is the percentage of documents approved by the user. The factor 4 indicates that from acceptance level larger then 25% the AF will have a contribution to increase the host ranking.

NH= Number of hosts searched.

RN= (number of relevant documents from that host)/(number of relevant documents from all hosts).

Also here, the factor NH indicates that if the percentage of documents approved from a given host is larger then  $1/NH$ , this section of the formula will contribute to the NHR.

4. For each host that was not found on the classified subjects list (such a host might have been referred to by a URL), calculate relevance using the formula of step 3, setting  $OHR=HRO$ ,  $HRO$  being the minimum host rank to be included in the CRD. The relevancy of the host is compared with the relevancy of the branch (of the classified subject list) to the query that evoked this host. A weighted relevance value may be calculated by a weighted average of the host relevancy and the classified subject list relevancy. If the weighted relevancy is above the predetermined value  $HRO$ , add this host to the CRD.

Based on existing standards such as the HTML, specific parameters of hosts may be automatically retrieved from the received data. For example, a received HTML page may be searched for specific expressions such as

<form action="A"> and <input type=submit value=B>

The URL address A and the value B may be tested for relevance to a search engine input page. This may be done by looking for words such as search, seek, explore, catalog and directory. If this page is identified as an input for a search engine, the parameters may be saved automatically in the CRD.

The validity of this page as an input to a search engine can be further tested by placing a query. The received page is then evaluated with the CSE for a search results typical characteristics such as:

- The expression *search w result\* w '</title>'*.
- The expression *match\**.
- The expression *camera\**.

The character \* indicates truncation boolean operator.

The page is ranked for being a result of a search process to validate the URL address as representing an input page to such a search engine.

It will be appreciated that the CRD updating, the formulas of step 3 and 4 and the HTML evaluation methods are described as an example and they do not limit the scope of the invention.

In yet another embodiment of the invention, a single CRD is shared by a group of users having the same interest (interest group - such as employees of a corporation acting in the field of DSPs - Digital Signal Processors). In this configuration, each user benefits from many searches made by his associates, and thus, an intensive updating and improvement of the shared CRD.

In another embodiment of the invention there is a CRD1 maintained and developed by specialists. CRD1 may have the advantage of covering a wide and dynamic source of information, such as the Internet. The maintenance of CRD1 will include new hosts that have been connected to the net, including their specific details such as their boolean search capabilities, available formats for query and a ranking of their value as information source. It may also include a special code for hosts that have been removed from the net, to avoid waste of time or even failures trying to contact these hosts. CRD1 is designed to be used by heterogeneous clients. As such, it is undesired that the update procedure of CRD1 will reflect the special needs for any individual client. This is a disadvantage for the individual client.

In order to benefit from the global update available in CRD1 and still enjoy the "personal" CRD that was updated to the client's need over many searches, CRD1 and CRD may both be used. The procedure of merging CRD1 and the client CRD can take place periodically, depends on how often the client is interested in it. The merge of CRD1 and the client CRD can be done in few ways.

In the first example it is assumed that the client's CRD was originally built by cutting a subgroup of subjects from CRD1. The client's CRD is then limited to the subjects of interest to that client and the client's CRD is a subgroup of CRD1:

1. Load CRD1 into the local station.
2. Compare each branch of client's CRD to the same CRD1 branch.
3. When new items are found in CRD1, like a new host in the Internet or even a new sub-subject, copy the relevant details to client's CRD.

In the second example, the client prefers to use the complete CRD1. The updating then done for all branches of CRD1. Through this procedure, the client has an updated CRD, modified to his own history of hosts ranking.

In another embodiment of the invention, The CRD may include not only names of hosts with databases which are bundled with their search engines. The CRD may also include URLs of locations that have no search engine that can be engaged to perform a search by a client. Such a URL may be useful, for example, if it refers to a URL/hypertext document that contains many URLs of other documents, all relevant to a specific subject. One such example is:

[http://web.mit.edu/afs/athena.mit.edu/org/i/invent/www/copyright\\_list.html](http://web.mit.edu/afs/athena.mit.edu/org/i/invent/www/copyright_list.html)

This URL refers to a document with the title:

"COPYRIGHTS, PATENTS, & RESOURCES FOR THE INVENTOR"

The document contains 18 titles with their URLs, and a short content description which is attached to most of them.

This URL is a very good candidate to become a part of a CRD, under classified subjects such as "Patents, Patents rights and Patents applications". Such a URL may become a part of the CRD by any of the CRD updating methods mentioned hereinabove such as: Highly ranked for providing many approved documents and manual update. It is appreciated that a CRD may contain only URLs. In such a case all the search process is done locally at the client's station.

A CRD containing both hosts and (database with associated search engine) and URLs of documents (usually hypertext documents) may take the following form:

<subject> [<host name>/<relevance ranking>, <URL>/<relevance ranking>, ...]

<URL>: The Internet address of that URL.

Considering relevance ranking, URLs may be treated indistinguishable from hosts and the priority for search will be based on relevance ranking only. This may be set as a user defined parameter so that priority for search may be any of:

- High relevance ranking first.
- First all hosts according to their relevance ranking, then all URLs according to their relevance ranking.
- First all URLs according to their relevance ranking, then all hosts according to their relevance ranking.

A CRD may contain reference only to URLs of documents. In such a case no search is evoked in the hosts. Information is retrieved based on the URL/hypertext method where the only search engine involved in the process is the CSE.

It will be appreciated that the CRD updating process may be completely automated, manually controlled or completely manual.

It will also be appreciated that the CRD updating method can be used to generate a completely new CRD. This is done according to the above mentioned step 4 of updating a CRD. The difference is that the process of steps 1-4 starts with a classified list that has no reference to information sources. By using the method of step 4, information sources are gradually filled in. The first host will be found by a URL/hypertext based search.

### 8. Creating and using a local database

In yet another embodiment of the invention, the documents approved by each of the members of an interest group are stored in a local server. As a result this server may become the most valuable information source for that interest group. Technically, this server is regarded as one more information providing host that naturally has a high rank. By setting the highest rank to this host, communication load on the net to other hosts may be reduced to a great extent and increase the search efficiency. It is appreciated that this method can be applied to any host on the net that is set to provide this kind of service to such an interest group.

In yet another embodiment of the invention, the local filtering of already received documents is done at the time interval starting after a query has been submitted and the time the host send the information found by the search engine of that host. By this method the local filtering and remote search are conducted generally in parallel. As a result the overall time required for this process is reduced.

The method described in this invention is easily applicable to net computers (NC), computers which are based on loading applications from the net instead of storing them on a disk. The application and reference data may be loaded as any other application. A personal CRD may be stored in a remote host that provides such personal services.

It is appreciated by those skilled in the art that the distribution of tasks and data bases among the components of the net (clients, local servers, remote servers) is extremely flexible and can be arranged in many forms. The embodiment described here is provided as an example and it is not limited to the described configuration.

Specific URLs mentioned hereinabove are relevant, as described, in the time that this description was written. It will be appreciated that at a later time the content of these URLs may be changed or they may be eliminated completely. These URLs are given as examples as they are at the time this description was written.

It will be appreciated that the embodiments described hereinabove are provided as examples and that the scope of the invention is not limited to this description.  
The scope of the invention is defined only by the claims hereinbelow.

What is claimed is:

1. A method for search of information in a digital information source, comprising:  
a first search engine, and;  
at least one digital information source, and;  
a second search engine, associated with said information source, and;  
a query composition interface for composing queries that are not restricted by the properties of said digital information source, and;  
the step of reducing the complexity of the boolean structure of the query to the level that is acceptable by the second search engine.
2. The method according to claim 1 including the step of filtering received data using said first search engine.
3. The method according to claim 1 including the step of ranking received data using said first search engine.
4. The method according to claim 2 and 3, where as any of the steps of filtering and ranking is active during the time interval between submitting a query and receiving the search results.
5. The method of reducing the complexity of the boolean structure of the query, including the step of:  
removing all boolean expressions preceded by the boolean operator 'NOT' and;  
generating a set of queries, each contains a single term of the terms not removed in the previous step.
6. The method according any of claims 1 through 3, including the step of consolidating information retrieved from multiple sources.
7. The method according any of claims 1 through 6, including the step of:  
maintaining a log list of reference to already retrieved information, and;  
using the log list to prevent multiple retrieval of the same information
8. The method according any of claims 1 through 6, including the steps of:  
indicating approved information, and;  
creating an information resource based on the accumulation of said approved information.
- 9B. A method for search of information in a digital information source, comprising:  
a first search engine, and;  
at least one digital information source, and;  
a second search engine, associated with said information source, and;  
a query composition interface for composing queries that are not restricted by the properties of said digital information source, and;  
the step of reducing the complexity of the boolean structure of the query to the level that is acceptable by the second search engine, and;  
a reference database of information sources.
10. The method according to claim 9 including:  
a classified subject list, and,  
the reference database of information sources is included in multiple levels of the classified subject list

11. The method according to any of claims 9 and 10 including at least one of the step of using search results to automatically add a new reference, of a new information source, to the reference database. and;  
the step of using search results to automatically modify an existing reference of an information source.
12. The method according to any of claims 9 through 11 including the step of updating a reference database according to another reference database.
13. The method according to any of claims 9 through 11 whereas the reference database is shared by at least two users
14. The method according to any of claims 9 through 11 including the step of merging at least two reference databases.
15. The method according to any of claim 1 whereas the reference database includes reference to information having a hypertext structure.
16. A method for search of information in a digital information source, comprising:  
a search engine, and;  
at least one digital information source, and;  
at least one hypertext reference document, and;  
a query composition interface for composing queries, and;  
the step of retrieving information using hypertext based automatic search, and;  
the step of filtering received data using said search engine.
17. A method for search of information in a digital information source, comprising:  
a search engine, and;  
at least one digital information source, and;  
a query composition interface for composing a query, and;  
the step of reducing the complexity of the boolean structure of the query to the level that is acceptable by the second search engine, and;  
a classified subjects list, and;  
a reference database of information sources that is associated with said classified subject list, and;  
the step of using said query and said search engine to automatically select information sources.
18. A method for search of information in a digital information source, comprising:  
a search engine, and;  
at least one digital information source, and;  
the step of reducing the complexity of the boolean structure of the query to the level that is acceptable by the second search engine, and;  
a classified subjects list, and;  
the step of creating a query by selecting a subject from the classified subject list.
19. The method of claim 18 and including the steps of:  
selecting all the terms in the branch preceding to the selected subject, including the selected subject itself, and;  
enclosing the keywords of each term in parenthesis and relating them by the boolean relation 'W' (adjacency). and;  
relating all the enclosed terms by the boolean operator 'AND'.
20. A method for search of information in a digital information source as described hereinabove.

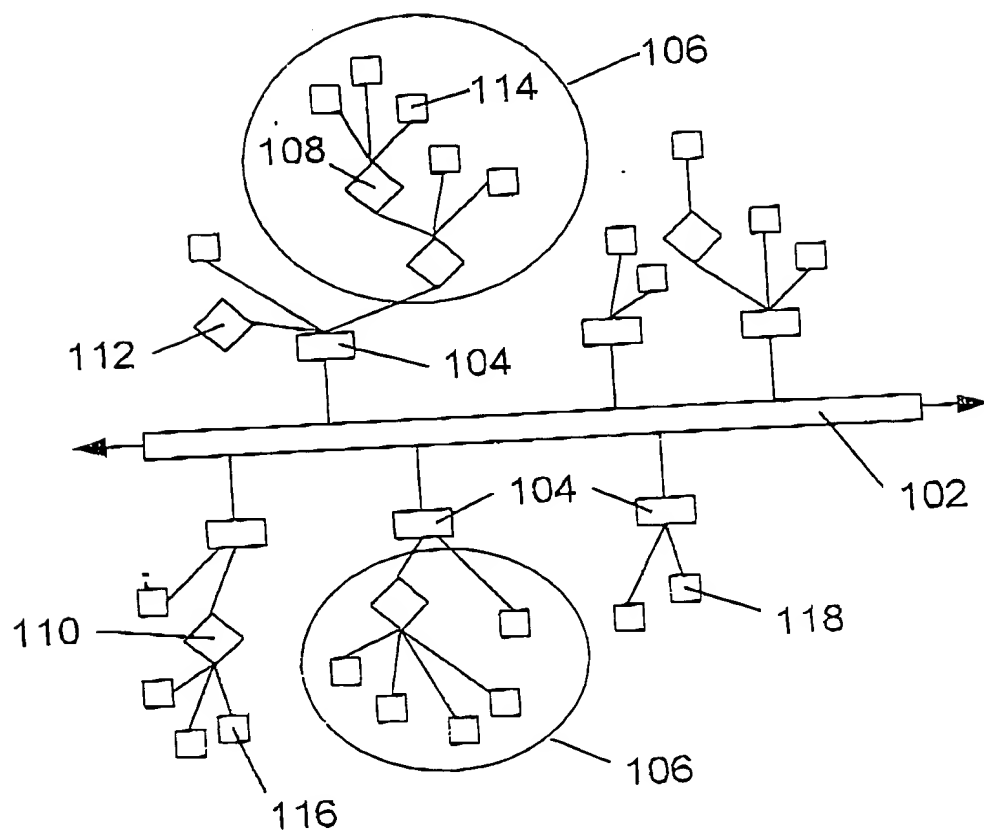


Figure 1

Haim Zvi Melman, Uzi Zurgil  
1(4)

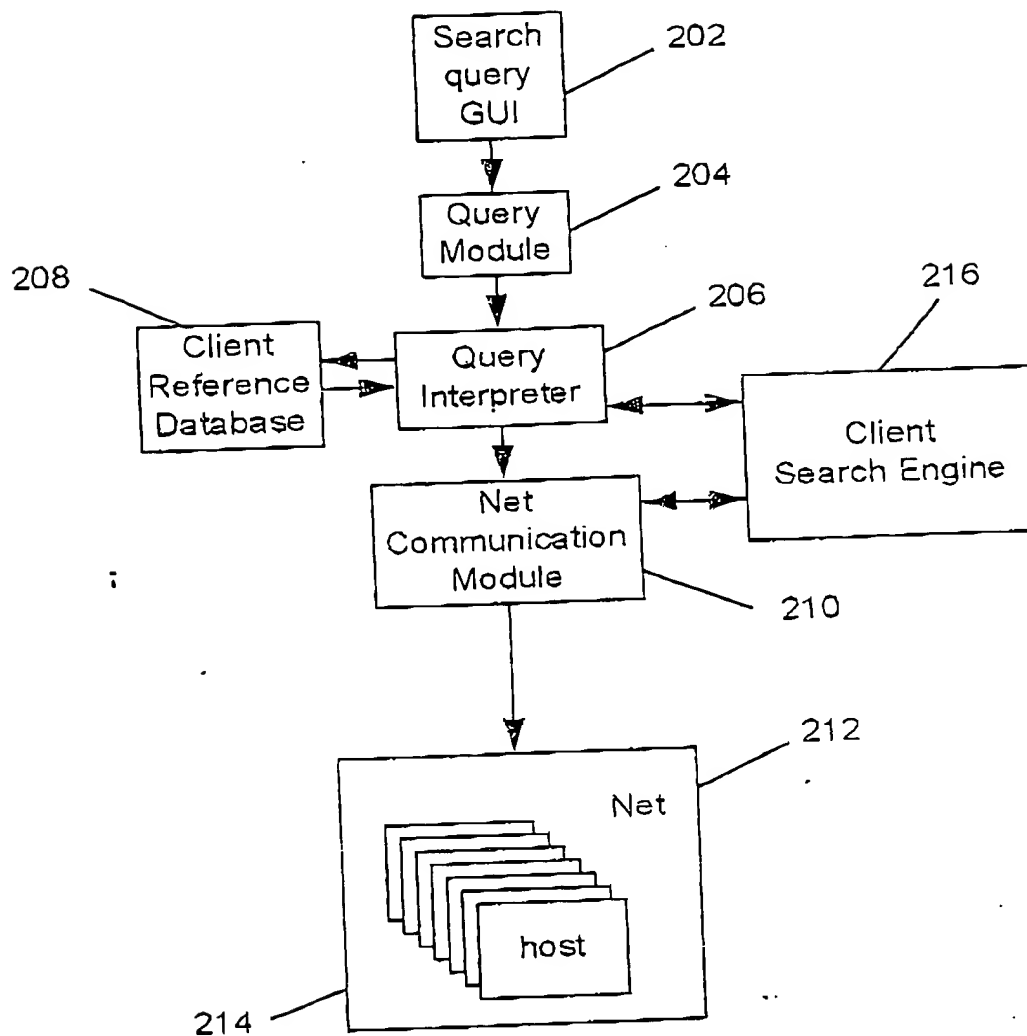


Figure 2

Haim Zvi Melman, Uzi Zurgil  
2(4)



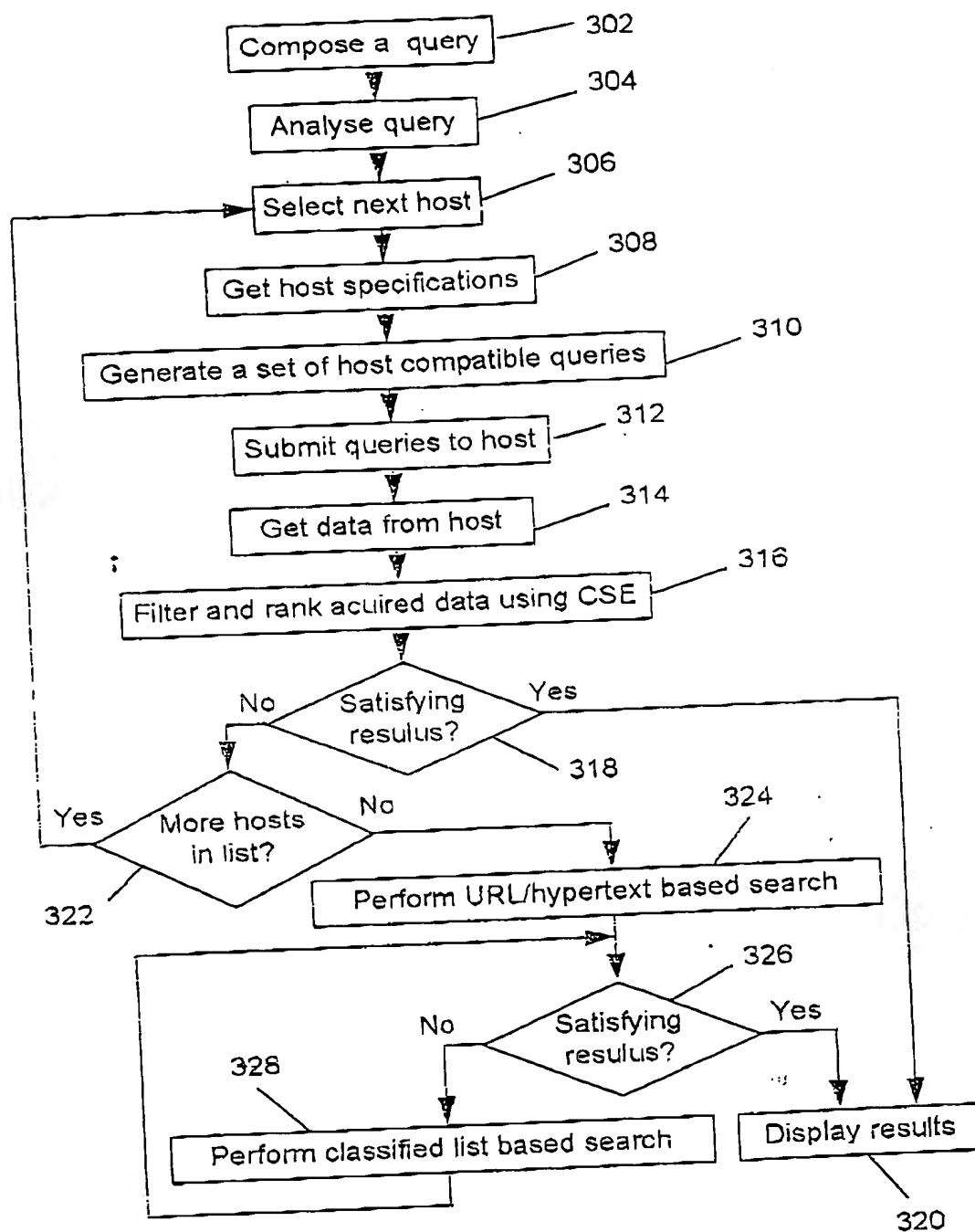


Figure 3

Haim Zvi Melman, Uzi Zurgil  
3(4)

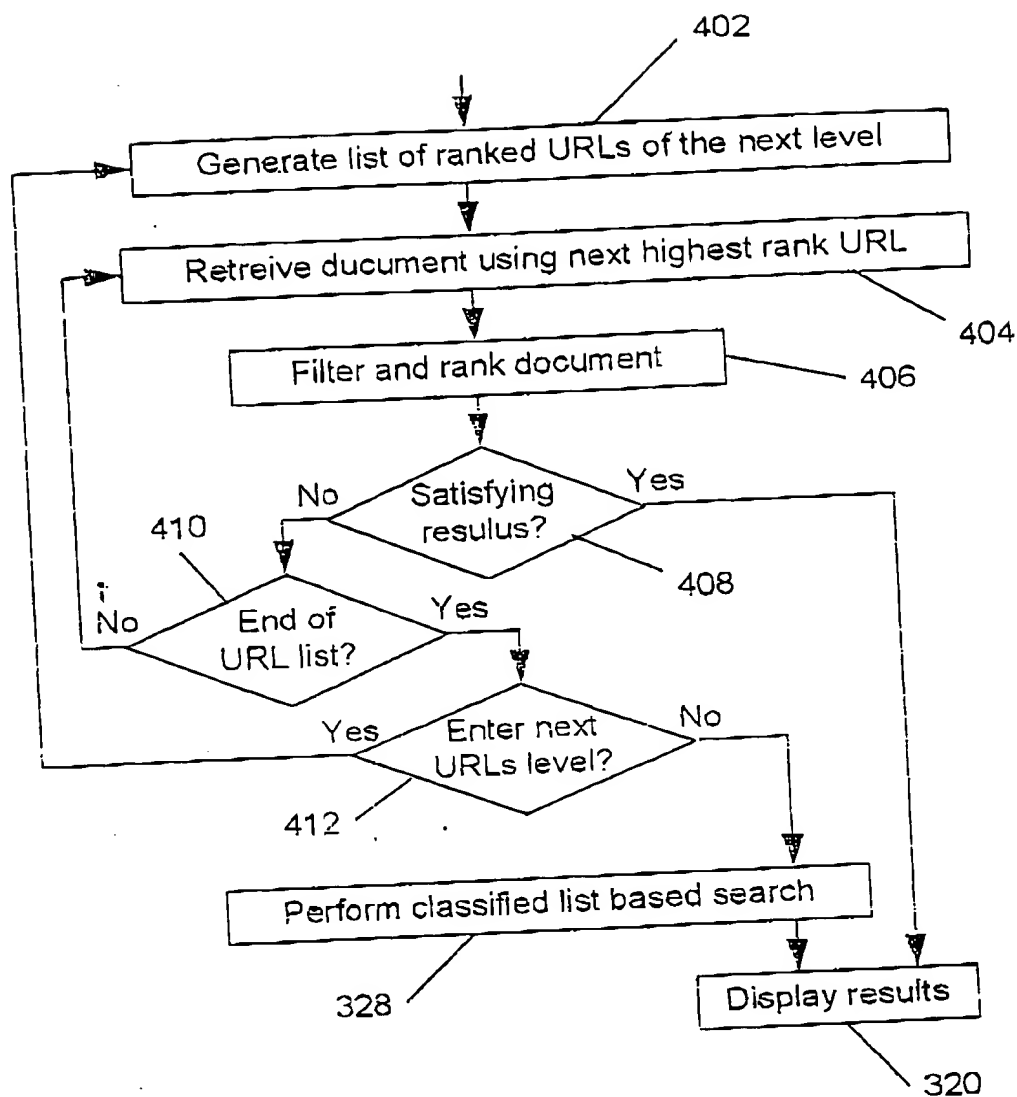


Figure 4

Haim Zvi Melman, Uzi Zurgil  
4(4)

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.